Return to
OISE

ARCHIVAL ONLY

DEALING WITH MISSING DATA IN MULTILEVEL ANALYSIS: A FIRST STEP

Bob Prosser
Ontario Institute for Studies in Education

Paper presented at the CSSE Conference,
Windsor, Ontario, June 5, 1988

ON05251

PROJECT REPORT

92-06913/

Progress Report: _____ Final Report: _____ Project Number: 3143

Small-Scale: xx Large-Scale: _____ Funding Year: _____

Transfer Grant: _____ (Priority: _____) OISE: _____

New: _____ Continuing: _____ Terminated: _____ Re-funded: _____

Project Title: A Comparison of Methods of Compensating for Missing Data in Estimation of Multilevel Model Parameters

Investigator(s): R. Traub

Budget Unit: MECA Date Rec'd in FS&R: _____

The purpose of this project was to conduct a Monte Carlo assessment of strategies to compensate for missing data in multilevel analysis. In general terms, the research objective was to determine the conditions under which application of strategy X led to less biased or more efficient estimates of multilevel model parameters than use of strategy Y. The budget was intended to cover costs associated with conducting the simulations on the University of Toronto's Cray X-MP computer.

During the spring of 1988, FORTRAN programs needed for the simulations were written and tested on OISE's (STS) VAX 11/750. (Listings of programs will be supplied on request.) Three methods of imputation for missing explanatory variable values -- mean, regression, and stochastic regression -- were selected for comparison in the initial pilot work. Several preliminary experiments were conducted to examine bias in and dispersion of parameter estimates obtained from randomly different data sets generated under various conditions. A paper reporting results of one experiment was presented at the CSSE conference. (Copy attached.)

The pilot studies to date have been conducted on the OISE VAX 8250 (the library computer) because it was (a) reasonably fast -- faster than a VAX 11/750 or a microcomputer, at least; (b) seldom used by anyone else; and (c) available at no (direct) cost for long periods of time. It seemed prudent to conserve resources while the "bugs" were being worked out in the study.

Refining the methodology for this study has been a lengthy process. As anticipated, repeated applications of the BILEVEL program (Goldstein, 1987) to provide the "raw material" for assessing estimate quality has been very time consuming. (Some of the experiments have taken many days of CPU time even on the VAX 8250.) A major problem in data generation was uncovered and corrected. Results from pilot runs have been encouraging. The strategies appear to have markedly different effects under different circumstances. Further, the experience is making it possible to design more useful experiments for the main study -- the range of sample sizes to be examined should be expanded, for example, and the number of

replications in each "cell" should probably be doubled (to improve estimate stability). The time has now come (albeit later than expected) to take advantage of the Cray's speed. (A free test was conducted on the Cray, and a 95% saving in CPU time was obtained on a typical run.)

The only expenditure of project funds so far has been to purchase a copy of HLM (Bryk, et al., 1986), an alternative computer program to fit multilevel models. (CSG will soon make this program available generally on STS via the STATS procedure.) It was considered important to compare the maximum likelihood estimates given by the BILEVEL program (using iterated generalized least squares) against those produced with a different algorithm (the EM algorithm).

Dealing with Missing Data in Multilevel Analysis: A First Step

Much individual behavior occurs within a group context, and it is an everyday observation that "people's behavior is influenced by their social environment" (van den Eeden & Hüttner, 1982). Multilevel research is concerned with understanding the interrelations among phenomena at different social levels. The term *level* is used here to suggest the notion of a hierarchical or nested membership relation among social units. An educational example of such a level structure is the grouping of students into schools. Within education again, one multilevel research area which has attracted considerable attention is the problem of determining the effects of school and student characteristics on individual student achievement (Aitkin & Longford, 1986).

Analysis of data from a hierarchical social system often involves combining information from units at two or more levels into a single (linear) model. Doing so makes it possible to include variation in macro-unit characteristics (school size, for example) and micro-unit explanatory variables (say, student ability) in accounting for differences among micro-units on some response measure. It is also possible to focus on within-group regression slopes as outcomes of interest and to model the variation (across macro-units) in these coefficients. When this is done, the micro-unit response of interest is regressed–in each macro-unit–on micro-unit explanatory variables. Variation in the rates of change of micro-units' "outputs" with their "inputs" is then accounted for in terms of differences in macro-unit characteristics. An example of the latter form of modelling is found in the work of Raudenbush and Bryk (1986) who examine, in part, the question of whether or not the relationship between individuals' achievement and socioeconomic status is stronger in public schools than in Catholic schools. Sector (Catholic/ public) was the level 2 variable whose effect on the achievement-SES slope was tested.

Although some multilevel research was conducted in the 1960's and the concept of slopes-as-outcomes dates back to at least the early 1970's, it is only in the past several years that attention has been directed to estimation and inference problems in multilevel modelling (De Leeuw & Kreft, 1986). Consideration has recently been given, for example, to the value of viewing within-group regression coefficients as random variables rather than as fixed parameters (Tate & Wongbundhit, 1983; Aitkin & Longford, 1986). Some early estimation approaches used ordinary least squares (OLS) to obtain coefficients for each group and OLS (or weighted least squares–see Hanushek, 1974) to relate these to group characteristics. As Bryk and Raudenbush (1986) have noted, because sampling variability in regression coefficients can be large in small groups, and because the precision of the coefficient estimates may vary across macro-units, use of OLS to examine the effects of group characteristics on slopes may reveal little. In the last three years, efficient procedures have been developed for two-level models (Longford, 1987; Bryk, Raudenbush, Seltzer, & Congdon, 1986) and for three-level models (Goldstein, 1986) which provide maximum likelihood estimates of parameters when random coefficients have a multivariate normal distribution.

Considerable work has been done to generalize the concept of multilevel modelling (Goldstein & McDonald, in press). Use of multilevel models can facilitate the analysis of survey data (Goldstein & Silver, in press), growth curve data (Bryk & Raudenbush, 1987), and multivariate and categorical data (Goldstein, 1987), for example. Despite the recent advances in the statistical aspects of multilevel research and the model's potential for widespread application, there are a number of practical issues of concern to users of multilevel analysis which remain to be explored in depth. One such issue is the selection of procedures to use in fitting multilevel models when data are incomplete.

The problem of not knowing the values of some variables for some units under study seems to be a common feature of social science research. Failure during statistical analysis to compensate adequately for the absence of the measurements–and, according to

Kalton (1983), such failure occurs frequently–results in bias and inefficiency in parameter estimation. In the words of Rubin (1987, p. 1): "missing values not only mean less efficient estimates because of the reduced size of the data base but also that standard complete-data methods cannot be immediately used to analyze the data. Moreover, possible biases exist because the respondents are systematically different from the nonrespondents; of particular concern, these biases are difficult to eliminate since the precise reasons for nonresponse are usually not known."

This paper presents some results from a Monte Carlo comparison–with respect to bias and efficiency–of compensation strategies (adapted from unilevel analysis) that could be used when estimating multilevel model parameters from incomplete data. The objective in this research is to provide data analysts with some reliable practical suggestions when explanatory variable values are missing at one or both levels in a two-level data set. The first section (of five) describes the multilevel model examined in the simulations, relating it to a broad class of two-level models and a possible substantive interpretation. The second section briefly outlines problems associated with missingness in the analysis of unilevel data, and discusses one empirical comparison of compensation strategies which might be used in ordinary regression analysis with incomplete data. The remaining sections discuss the methodology, results, and conclusions.

1 A SIMPLE CASE OF A GENERAL RANDOM COEFFICIENT REGRESSION MODEL

De Leeuw and Kreft (1986) have described a two-level model which will be referred to here as a general random coefficient regression (GRCR) model. Models which are more general still have been described by Goldstein and McDonald (in press), but the GRCR version serves our purposes of illustrating the slopes-as-outcomes concept in multilevel modelling and providing a "parent" for the simple exemplar used in this study.

1.1 Structure and Assumptions of the GRCR Model

In this discussion, level 1 observational units will be called *cases*; these are contained in level 2 units called *groups*. For simplicity, each case belongs to only one group. Suppose there are G groups containing a total of n cases, with n_j cases in group j. The two-level model for group j can be written as

$$\mathbf{y}_j = \mathbf{Z}_j \underline{\eta}_j + \underline{\epsilon}_j \tag{1}$$

where $\mathbf{y}_j$ is an n_j-vector of response measurements; $\mathbf{Z}_j$ is an $n_j \times P$ matrix consisting of (fixed) values of level 1 explanatory variables; $\underline{\eta}_j$ is a P-vector of within-group random regression coefficients; and $\underline{\epsilon}_j$ is an n_j-vector of case-level random disturbances. The regression coefficients are random in the sense that they can be considered "random drawings from a probability distribution" (Judge et al., 1985, p. 544); fixed coefficients would be a special case. It is assumed that random coefficients are not correlated with the level 1 explanatory variables.

Combining the models for the G groups gives

$$\mathbf{y} = \mathbf{Z}\underline{\eta} + \underline{\epsilon} \tag{2}$$

where $\mathbf{y} = [\mathbf{y}_1', \ldots, \mathbf{y}_G']'$; $\mathbf{Z} = \sum_j^+ \mathbf{Z}_j$; $\underline{\eta} = [\underline{\eta}_1', \ldots, \underline{\eta}_G']'$; and $\underline{\epsilon} = [\underline{\epsilon}_1', \ldots, \underline{\epsilon}_G']'$. The symbol $\sum_j^+$ indicates the direct sum operation. The assumptions made about $\underline{\epsilon}$ are as follows: (a) $\mathcal{E}(\underline{\epsilon}) = \mathbf{0}$; and (b) $\mathcal{D}(\underline{\epsilon}) = \sum_j^+ \mathbf{\Phi}_j$. De Leeuw and Kreft (1986) make the simplifying assumption that $\mathbf{\Phi}_j = \sigma_{\epsilon_j}^2 \mathbf{I}_{n_j}$.

Each coefficient η_{jp} $(j = 1, \ldots, G; p = 1, \ldots P)$ can be modelled as a linear combination of group-level explanatory variables:

$$\eta_{jp} = \mathbf{w}_{jp}' \underline{\beta}_p + \upsilon_{jp} \tag{3}$$

where vector $\mathbf{w}_{jp}$ consists of q_p fixed level 2 measurements; $\underline{\beta}_p$ is a q_p-vector of fixed parameters; and υ_{jp} is a group level random disturbance. The random disturbances can be

considered to express effects of unspecified variables on the η_{jp}s. The variance of υ_p represents "true" variation in the pth coefficient that is not accounted for by the explanatory variables in the $\mathbf{w}_p$s. (In the simplest situation, η_{jp} is simply $\beta_p + \upsilon_{jp}$, i.e., the random coefficient is modelled as a population mean value and a random deviation/ effect for each group.)

Defining $P \times Q$ matrix $\mathbf{W}_j$ as $\sum_j^+ \mathbf{w}'_{jp}$, (where $Q = \sum_{p=1}^{P} q_p$) the model for group j coefficients can be expressed as

$$\underline{\eta}_j = \mathbf{W}_j\underline{\beta} + \underline{\upsilon}_j \tag{4}$$

with Q-vector $\underline{\beta}$ given by $[\underline{\beta}'_1, \ldots, \underline{\beta}'_P]'$. Assumptions about disturbance P-vector $\underline{\upsilon}_j$ are as follows: (a) $\mathcal{E}(\underline{\upsilon}_j) = \mathbf{0}$; (b) $\mathcal{E}(\underline{\upsilon}_j\underline{\upsilon}'_{j'}) = \mathbf{0}$ if $j \neq j'$; and (c) $\mathcal{E}(\underline{\upsilon}_j\underline{\upsilon}'_{j'}) = \mathbf{\Upsilon}_j$ if $j = j'$. It is also assumed that every element of $\underline{\epsilon}_j$ is distributed independently of every element in $\underline{\upsilon}_j$ i.e. $\mathcal{C}(\underline{\epsilon}_j, \underline{\upsilon}_j) = \mathbf{0}$.

Stacking the G models for random coefficients gives

$$\underline{\eta} = \mathbf{W}\underline{\beta} + \underline{\upsilon}. \tag{5}$$

Here, $\underline{\eta} = [\underline{\eta}'_1, \ldots, \underline{\eta}'_G]'$; $\mathbf{W} = [\mathbf{W}'_1, \ldots, \mathbf{W}'_G]'$; $\underline{\beta} = [\underline{\beta}'_1, \ldots, \underline{\beta}'_G]'$; and $\underline{\upsilon} = [\underline{\upsilon}'_1, \ldots, \underline{\upsilon}'_G]'$. The model for $\underline{\eta}$ can be incorporated into equation 2 giving:

$$\begin{aligned} \mathbf{y} &= \mathbf{ZW}\underline{\beta} + \mathbf{Z}\underline{\upsilon} + \underline{\epsilon} \\ &= \mathbf{X}\underline{\beta} + \underline{\nu} \end{aligned} \tag{6}$$

where $\mathbf{X} = \mathbf{ZW}$ and $\underline{\nu} = \mathbf{Z}\underline{\upsilon} + \underline{\epsilon}$. Vector $\mathbf{X}\underline{\beta}$ is called the fixed part of the model, and $\underline{\nu}$ is called the random part. Earlier assumptions about the expectation and dispersion of $\underline{\upsilon}$ and $\underline{\epsilon}$ lead to the following statements:

$$\mathcal{E}(\underline{\nu}) = \mathbf{0}; \quad \text{and}$$

$$\mathcal{D}(\underline{\nu}) = \mathbf{\Sigma} = \sum_j^+ \mathbf{\Sigma}_j$$

where $\mathbf{\Sigma}_j = \mathbf{Z}_j\mathbf{\Upsilon}_j\mathbf{Z}'_j + \mathbf{\Phi}_j$. Often two simplifications of the model are introduced:

(a) $\mathbf{\Phi}_j = \sigma^2_{\epsilon_j}\mathbf{I}_{n_j} = \sigma^2_{\epsilon}\mathbf{I}_{n_j}$; and

(b) $\mathbf{\Upsilon}_j = \mathbf{\Upsilon}$ $(j = 1, \ldots, G)$.

Here, both have been adopted. When maximum likelihood estimation is performed, it is usually assumed that the disturbances at each level have multivariate normal distributions.

The GRCR model parameters to be estimated are as follows: $\underline{\beta}$, Υ, and σ^2_ϵ. In some forms of multilevel research–school effectiveness research is a prime example–level 2 residuals, $[\underline{v}'_1, \ldots, \underline{v}'_G]'$, are estimated for substantive reasons, not simply for diagnostic purposes (Aitkin & Longford, 1986).

1.2 The Exemplar Used in the Simulations

The model used in the initial simulations had the following structure:

$$\begin{aligned} y_{ij} &= [\,1, z_{1ij}\,] \begin{bmatrix} \eta_{0j} \\ \eta_{1j} \end{bmatrix} + \epsilon_{ij} \\ &= [\,1, z_{1ij}\,] \left(\begin{bmatrix} 1 & w_{1j} & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \beta_{00} \\ \beta_{01} \\ \beta_{10} \end{bmatrix} + \begin{bmatrix} v_{0j} \\ v_{1j} \end{bmatrix} \right) + \epsilon_{ij} \end{aligned} \tag{7}$$

It was selected for its simplicity, having one explanatory variable at each level, three fixed parameters (β_{00}, β_{01}, β_{10}), and four random parameters (σ^2_0, σ_{01}, σ^2_1, σ^2_ϵ).

For illustrative purposes, a (grossly oversimplified) situation from school effects research will be described for the model given by equation 7. The key features of this situation are as follows:

(a) a response variable, y_{ij}, (say achievement of student i in school j) is a function of a case-level explanatory variable, z_{ij} (say socioeconomic status of the family of student ij);

(b) the mean level of achievement for school j, expressed by η_{0j}, is linearly related to a measure of the school's level of resources, w_j;

(c) the relationship between individual achievement and socioeconomic status, η_{1j}, varies across schools, but this variation is not viewed as being systematically related to school resources or any other school-level variable that has been measured.

The parameters to be estimated are as follows:

(a) the mean level of achievement in the schools, β_{00};

(b) the mean rate of change of achievement with school resources, β_{01};

(c) the mean rate of change of achievement with socioeconomic status, β_{10};

(d) the variability in schools' average achievement levels, not accounted for by school resources differences, σ_0^2

(e) the variability in schools' rates of change of student achievement with socioeconomic status, σ_1^2; and

(f) the conditional covariance of schools' average achievement and rate of change, σ_{01}. (This covariance is conditional on resources.)

(g) the level 1 residual variance, σ_ϵ^2.

In this study, the focus is on the four random parameters since little work appears to have been done concerning treatment of *partially* missing data in variance component estimation. (There is now a fairly extensive literature on mixed model estimation from unbalanced experimental designs–see, for example, Searle, 1987.)

2 MISSING DATA

A simple unilevel univariate "case study" discussed by Kalton (1983) and Rubin (1987) will be outlined first to illustrate (i) difficulties to be faced in estimating parameters when data are incomplete and (ii) several of the many methods which have been proposed for use in such analysis. Assumptions about the nature of the missing data and strategies to partially compensate for incomplete data in unilevel analysis will then be discussed further. Missingness in multilevel research will then be briefly examined and suggestions for compensation strategies will be given.

Consider a population of size N (where N is very large) in which each element has a (true) value on some interval variable Y. The population mean value for this variable

is $\overline{Y}$ and the population variance is S^2. The researcher's task is to obtain point estimates of these quantities from a simple random sample of size n. In the population, N_r elements, say, would provide values (*respond*) if sampled and N_m $(= N - N_r)$ would not. Further suppose that the population mean for the N_r responders is $\overline{Y}_r$ and that the population mean for the nonresponders is $\overline{Y}_m$; the respective variances are S_r^2 and S_m^2. (Clearly $\overline{Y} = \frac{N_r}{N}\overline{Y}_r + \frac{N_m}{N}\overline{Y}_m$.) In the particular sample drawn, n_r units supply measurements and n_m do not; the mean for the n_r respondents is say $\overline{y}_r$. At this point, nothing has been assumed about the nature of the responders' Y values vis-a-vis those of the nonresponders.

The researcher has a choice among strategies to partially compensate for the data deficiency. Three will be discussed here. A "do nothing" (or *complete data* or *listwise deletion*) approach in this example would be to use $\overline{y}_r$ as the estimator of $\overline{Y}$ and s_r^2 as the estimator of S^2. The bias of this estimate of $\overline{Y}$ is $\mathcal{E}(\overline{y}_r - \overline{Y}) = \overline{Y}_r - \overline{Y} = \frac{N_m}{N}(\overline{Y}_r - \overline{Y}_m)$. In words, the bias is the product of the (population) missingness fraction and the difference between responder and nonresponder means. It can be shown (Kalton, 1983, p. 9) that for large N, N_r, and N_m, the bias involved in using s_r^2 to estimate S^2 is approximately $\frac{N_m}{N}(S_r^2 - S_m^2) - \frac{N_r N_m}{N^2}(\overline{Y}_r - \overline{Y}_m)^2$. Under either of the following assumptions

(a) $\overline{Y}_r = \overline{Y}_m$ and $S_r^2 = S_m^2$ or

(b) the responder and nonresponder populations are (very large) random samples from the same superpopulation

$\overline{y}_r$ and s_r^2 are unbiased estimators. Inferences about $\overline{Y}$ can then be made using the fact that $\overline{y}_r \dot{\sim} \mathcal{N}(\overline{Y}, s_r^2(n_r^{-1} - N^{-1}))$.

A second strategy would be to impute the best predicted value for each of the unobserved values, treating the fabricated "data" as real. In the absence of additional information, this amounts to assigning the value $\overline{y}_r$ to each of the n_m nonrespondents in the sample. (This strategy will be called *mean value imputation* and abbreviated MV.) Clearly $\overline{y}_{MV} = \overline{y}_r$ and $\hat{S}_{MV}^2 = s_r^2\frac{(n_r-1)}{(n-1)}$. Conditional on the value of n_r and assuming (a)

or (b) above, the bias for the estimator $\hat{S}^2_{\text{MV}}$ is $\frac{-n_m}{(n-1)}S^2$. Inference about $\overline{Y}$ based upon $\hat{S}^2_{\text{MV}}(n^{-1} - N^{-1})$ would be incorrect because the sample size used is inflated and because $\hat{S}^2_{\text{MV}}$ is too small.

The final strategy to be considered in this case study is *random imputation* (RI): the sample's nonrespondents are assigned Y values obtained from respondents by random sampling with replacement. If $\overline{y}_{\text{imp}}$ and s^2_{imp} are the mean and variance, respectively, of the n_m imputed values (when a particular sample has been chosen), $\overline{y}_{\text{RI}}$ is given by $\frac{n_r}{n}\overline{y}_r + \frac{n_m}{n}\overline{y}_{\text{imp}}$ and $\hat{S}^2_{\text{RI}}$ is computed using the following formula:

$$\hat{S}^2_{\text{RI}} = \frac{(n_r - 1)s^2_r + (n_m - 1)s^2_{\text{imp}} + n_r n_m n^{-1}(\overline{y}_r - \overline{y}_{\text{imp}})^2}{n - 1}$$

When (b) holds, $\mathcal{E}(\overline{y}_{\text{RI}}) = \overline{Y}$ and, conditional on the observed data, $\mathcal{E}(\hat{S}^2_{\text{RI}}) = s^2_r(1 - \frac{1}{n_r})(1 + \frac{n_r}{n^2 - n}) \approx s^2_r$. The variance of $\overline{y}_{\text{RI}}$ is $\frac{n_m}{n^2}(1 - \frac{1}{n_r})s^2_r$, and this is larger than the variance of the listwise deletion estimator $\overline{y}_r$ by a factor of $1 + \frac{(n_r - 1)(1 - n_r/n)}{n(1 - n_r/N)}$, thus the random imputation strategy is less efficient than the listwise deletion strategy (Rubin, 1987).

Many techniques have been devised for use in *unilevel* multivariate analyses to compensate, at least partially, for data incompleteness. The most commonly used methods involve the assumption that data are *missing at random* (MAR) or *missing completely at random* (MCAR). It is important to recognize that varying assumptions about the nature of the missing observations are implicit in the different treatments and that the importance of the assumptions varies with the type of analysis to be conducted (Little & Rubin, 1987). The "problem" of missing data is, in a sense, living with the consequences of the choices we make.* A simple example adapted from Little and Rubin (1987) can provide an informal description of these concepts. One can think of the elements of a matrix of data from population members as being governed by a *responding mechanism*–a stochastic process

* Rubin(1987) has described a technique he calls *multiple imputation* which allows one to test the sensitivity of one's conclusions to alterations of assumptions about missingness.

determining which sample units provide measurements on a given variable. Suppose that in a population, all persons will tell their age but only some will tell their income. If the probability that a sampled person tells his/ her income does not depend on either his/ her age or income, the data are MCAR. The persons who supply both age and income values are, in effect, a random subsample of the selected sample. It is possible that the probability of obtaining an income response depends on age but not on income; in this case the data is MAR. When the responding mechanism for income is a function of income, the data are neither MCAR or MAR.

Marini, Olsen, and Rubin (1980) have discussed the common compensation strategies in an framework using four categories.

(a) In some studies, cases which are missing values on one or more variables are omitted from all analyses. In this listwise deletion approach, analysis proceeds as if a sample of size n_r had been selected rather than a sample of the intended size, n ($n_r < n$).

(b) One way to make more complete use of the available information is to use a *pairwise deletion* strategy. In computating the sample mean and variance for a variable X, all cases in the sample with responses on X are used, and in calculating the covariance of any given pair of variables, all cases having values on both are involved.

(c) A third strategy–actually several families of approaches–is "filling in the gaps" or *imputation*. Imputation may be "deterministic" or "stochastic" depending upon whether or not the number of possible values which a technique could assign to a given unit on a given variable is one or more than one (Santos, 1981). Among the families of approaches are the following:

- assignment of the sample mean for a variable to those cases missing real values for the variable. A refinement of this deterministic method is to split the sample into subgroups and assign subgroup means.
- random donation of values within subgroups (the subgroups having been formed on the basis of variables which tend to distinguish nonrespondents

from respondents). There are many ways in which donation is done in practice. One is to assign to a nonrespondent the value of the variable in question from the closest previous respondent his subsection of the data file.

- model-based imputation. Assumptions are made about the nature of the nonrespondents' values on variables of interest and a model e.g. a regression model is used to predict unobserved values for each case from observed information. In stochastic versions of these procedures, random residuals are added to predicted values.

(d) Maximum-likelihood procedures may be used to estimate parameters of interest directly, i.e., without calculating values to impute. Complete and incomplete data vectors are combined in a likelihood function for parameters of interest, commonly based on an assumption of multivariate normality. The missing values are treated as additional unknown quantities (along with the parameters) and the known data are treated as fixed. In general, the function is maximized iteratively, although for special patterns of missingness (e.g. monotone) noniterative computations of estimates can be performed (Marini et al., 1980). The increasingly ubiquitous EM algorithm (Dempster, Laird, & Rubin, 1977) is an example of a technique in this category.

Use of methods in the first three categories are appropriate if the data are MCAR (or MCAR within subgroups for the more complex methods of the third type.) Likelihood methods can be used when data are MAR.

In many situations, one cannot assume that missing data are MCAR or MAR. Data are sometimes missing for understandable reasons: sample members with values of a variable below a particular threshold may be less likely than others to supply information on that variable, for example. People with little formal education may be less likely than

others to report their level of education on a questionnaire. Values of one variable may tend to be unavailable primarily when values of another variable are missing. When data are neither MCAR or MAR, it is sometimes possible to model the stochastic response process–a distribution function, with unknown parameter values, for an indicator of the occurence of response–and improve on estimates that would be obtained if the mechanism were ignored.

Simulation studies have been conducted to compare effects of different compensation strategies under a variety of conditions when fitting unilevel regression models to incomplete data (Beale & Little, 1975; Little, 1979; Santos, 1981; Basilevsky, Hum, Sabourin, & Anderson, 1985). Analytical comparisons, particularly for regression involving more than one explanatory variable, are difficult and rare. Basilevsky et al. (1985), in the only study analyzed as a simulation *experiment*, compared nine strategies from the four categories noted above. Factors examined for their influence on strategy effects included level of R^2, multicollinearity, sample size, and fraction of values missing (completely at random). Three criteria including mean square error for the coefficients were used in assessing the effects of the factors. The researchers concluded that no strategy was uniformly best, and made recommendations about the strategy-of-choice in different circumstances. More specifically, they found that strategies like listwise deletion and mean imputation performed well when only 10% of the data were missing and that "for 30% of data missing, for both large and small sample sizes, the listwise and pairwise deletions as well as the mean value replacement are best." Part of the practical significance of the research lies in the fact that often strategies that are simpler and cheaper to implement perform as well as or better than complex imputation or maximum likelihood strategies.

In a study of a hierarchical social system, data may be missing at any or all levels, and it seems reasonable to suggest that (sometimes, at least) the response mechanisms at the different levels may operate independently. In a two-level study of achievement

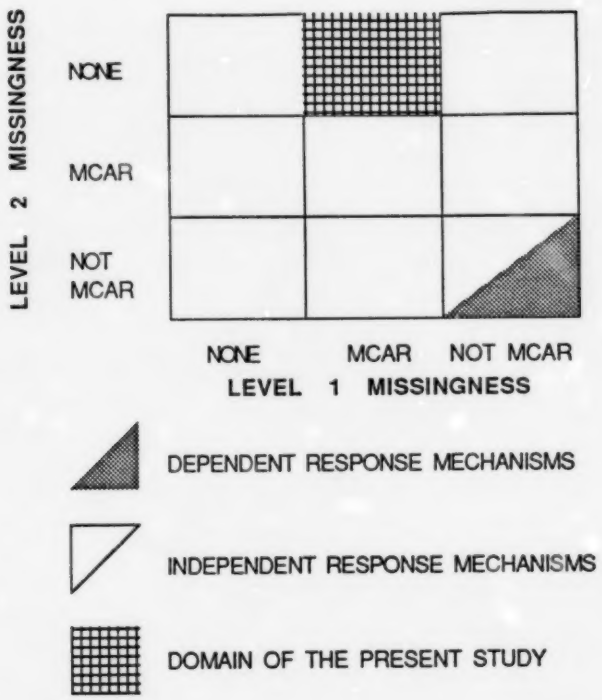


FIGURE 1: Possible Forms of Incompleteness in Multilevel Data

in which students supply test scores and teachers answer questionnaires concerning classroom practices, there may be no relationship between a teacher's propensity to state the percentage of instructional time that was devoted to a particular topic during the term and a student's likelihood of being absent on the day of a pretest. Further, data may be MCAR at one level, all levels, or at no level. Figure 1 displays the possibilities for incompleteness. When data are MCAR at some level, it may well be that application of simple compensation techniques at that level will, under some conditions, yield suitable estimates for model parameters.

To date, no comparative research on treatment of missing values in multilevel analysis has been reported. This attempt to shed some light on the practical problem of estimating two-level model parameters when values of explanatory variables are MCAR

for level 1 and level 2 units is patterned–to a degree–on the study by Basilevsky and his colleagues. The preliminary work has focussed on missingness at the lower level.

3 METHODOLOGY

A simulation experiment comprised of 5 blocks of 50 replications was conducted to examine the effects of sample size/ number of groups–abbreviated G–on random parameter estimate quality when each of three imputation methods was applied to incomplete data sets. In this section, four aspects of the procedure will be summarized: data generation and estimation; imputation methods; experimental design; and assessment of estimation quality.

3.1 Data Generation and Estimation

The model presented in equation (7) was used with the following arbitrary values assigned to the parameters in all blocks of replications: $\beta_{00} = 20$; $\beta_{01} = 10$; $\beta_{10} = 1$; $\sigma^2_{v_0} = 10$; $\sigma_{v_0 v_1} = 2$; $\sigma^2_{v_1} = 1$; and $\sigma^2_{\epsilon} = 20$.

Each replication consisted of five steps: (a) generating a set of $[Y, W, Z]$ vectors; (b) estimating *complete data* values of the random parameters; (c) designating some Z values to be missing in each group; (d) imputing values for the missing Zs; and (e) obtaining new estimates of the random parameters. Steps (d) and (e) were performed for each imputation method. Because the same complete/ incomplete data vectors were used for each fitting of the model on a replication, method was a within-subjects factor.

Estimation of the random parameters is conditional on fixed values of the explanatory variables, so a "standard" set of Z and W values was used for every replicate in a block. Level 1 and level 2 disturbance sets–these varied from one replication to the next–were obtained using GGNSM (IMSL, 1980), a generator of multivariate normal deviates. The situation simulated can be viewed as one in which repeated measurements (here, 50)

are made on the cases in a set of groups. Explanatory variable values at both levels do not vary from one measurement occasion to the next, but the sizes of the random coefficients do. On each occasion, in a manner completely unrelated to the value of any measured or unmeasured variable, some cases in each group fail to provide their values of the level 1 explanatory variable.

An iterative generalized least squares (IGLS) algorithm was used in obtaining maximum likelihood estimates of the parameters. Version 12-87 of Goldstein's (1986) BILEVEL program was adapted for the experiment: the data generation procedure replaced the data read-in and the write-out was modified to produce estimate vectors that could be analyzed with standard statistical packages. All simulations were performed on the O.I.S.E. VAX 8250.

3.2 Imputation Methods

Three methods for replacing missing Zs were investigated:

(a) mean value (MV);

(b) regression (R); and

(c) stochastic regression (SR).

Imputation was conducted in each group individually. In MV imputation, $\overline{z}_{g.}$ was assigned to each of the n_m receiver cases in group g ($g = 1, \ldots, G$). Regression imputation involved two steps: (a) building a regression function in a group using cases with complete data ($\hat{z}_{gi} = \hat{a}_g + \hat{b}_g y_{gi}$); and (b) using the Y values of the incomplete cases to estimate their Zs. Two additional steps were added to this process to arrive at SR imputed values: (c) obtaining residual estimates $\hat{e}_{gi}$ ($= z_{gi} - \hat{z}_{gi}$) from the complete cases; and (d) randomly selecting n_m of these and adding them to the R imputed values.

3.3 Experimental Design

To examine the effects of G on estimate quality, a two factor (G by imputation method) repeated measures design was used. Five levels of the "between-subjects" factor, G, were used:

(a) 15 groups of 30 cases;

(b) 30 groups of 30 cases;

(c) 50 groups of 30 cases;

(d) 75 groups of 20 cases; and

(e) 100 groups of 15 cases.

In each of the five blocks of 50 replications, the percentage of Z values missing in each group was 20.

3.4 Assessment of Estimate Quality

Indicators of estimate bias, precision, and accuracy were computed for each of the four parameters under each G level and imputation method combination. The primary quantities used in the computation of these indicators were *relative deviates*. If θ is the true value of a particular parameter and $\hat{\theta}_{ckm}$ is the estimate of the parameter from the cth replication under the kth set of conditions and the mth imputation method, the corresponding relative deviate $\hat{\vartheta}_{ckm}$ is given by:

$$\hat{\vartheta}_{ckm} = \frac{\hat{\theta}_{ckm} - \theta}{\theta} \tag{8}$$

For each set of conditions and imputation method, the mean and standard deviation of the 50 relative deviates were computed. The mean served as an index of bias, and the standard deviation was used as a measure of accuracy. In addition, the root of the mean of the squared deviates was computed as an indicator of precision.

4 RESULTS

The findings for the estimates of $\sigma^2_{v_0}$, $\sigma_{v_0 v_1}$, $\sigma^2_{v_1}$, and σ^2_ϵ will be discussed in turn. The bias, precision, and accuracy values for each type of estimate–the basis for the graphs and analyses presented here–are contained in Table 1. One-factor repeated measures analyses of variance were conducted to test for method differences in bias, precision, and accuracy. (In these ANOVAs, G was the "subjects" factor.) Table 2 presents the F ratios and significance probabilities* for these tests.

4.1 Estimates of Intercept Disturbance Variance

The bias index values in column 1 of Table 1 show that intercept disturbance was underestimated at each G level when the model was fitted to complete data, and when each imputation method was applied. Bias was most severe when the number of groups was smallest–at least 20% for each imputation method–and improved to approximately 10% when the number of groups was 75, but it did not change monotonically with G. When the number of groups was 100 with a group size of 15, bias was more negative than when the number of groups was 75 with a group size of 20. Further investigation is needed concerning the relative impact on bias of group size and number of groups.

Figure 2a displays the bias values obtained for each G level and imputation method. The ANOVA test provided evidence of method differences in bias ($F_{2,8} = 13.9; p = 0.015$). For each level of the G factor, the method which produced the greatest downward bias was mean value (MV) imputation. Use of regression (R) and stochastic regression (SR) imputation resulted in nearly equal levels of bias.

The precision values for each type of estimate are displayed for each G level and method combination in Figure 3a. For all imputation methods, precision improved

* Because tests of within-subjects factors tend to be liberal, the significance probabilities for adjusted F tests using Huynh and Feldt's (1976) ε estimate were examined.

		σ_0^2				σ_{01}				σ_1^2				σ_ε^2			
	G	CD*	MV	SR	R	CD	MV	SR	R	CD	MV	SR	R	CD	MV	SR	R
BIAS	15	-.20	-.24	-.22	-.22	-.03	-.10	-.04	.04	-.03	-.06	.08	.26	.01	.15	.-00	-.06
	30	-.13	-.13	-.13	-.12	-.03	-.02	-.02	.07	-.01	-.01	.06	.26	.00	.11	-.02	-.07
	50	-.11	-.12	-.11	-.11	-.02	-.01	.01	.09	-.02	-.02	.06	.27	.00	.11	-.02	-.07
	75	-.12	-.12	-.11	-.10	-.01	-.01	.01	.11	-.01	-.05	.10	.31	.00	.10	-.02	-.07
	100	-.16	-.17	-.15	-.14	-.03	-.04	.01	.08	-.03	-.09	.12	.32	.00	.12	-.03	-.08
		CD	MV	SR	R	CD	MV	SR	R	CD	MV	SR	R	CD	MV	SR	R
PRECISION	15	.40	.42	.40	.40	.51	.50	.52	.55	.40	.44	.45	57	.08	.17	.08	.10
	30	.29	.29	.28	.28	.35	.37	.36	.41	.31	.34	.34	.48	.05	.12	.06	.09
	50	.20	.20	.19	.19	.26	.28	.27	.31	.24	.25	.27	.40	.04	.11	.04	.08
	75	.19	.19	.18	.18	.22	.24	.24	.28	.21	.23	.25	.41	.04	.11	.04	.08
	100	.20	.21	.19	.19	.19	.20	.20	.22	.16	.20	.22	.38	.03	.12	.04	.08
		CD	MV	SR	R	CD	MV	SR	R	CD	MV	SR	R	CD	MV	SR	R
ACCURACY	15	.35	.34	.34	.34	.51	.50	.53	.56	.40	.43	.45	.51	.08	.09	.08	.08
	30	.26	.26	.26	.26	.36	.37	.36	.40	.31	.34	.34	.40	.05	.06	.05	.05
	50	.16	.16	.16	.16	.27	.28	.28	.30	.24	.25	.26	.30	.04	.04	.04	.04
	75	.14	.14	.14	.14	.22	.24	.24	.26	.21	.22	.24	.27	.04	.05	.04	.04
	100	.12	.12	.12	.12	.19	.19	.21	.21	.16	.17	.19	.21	.03	.03	.03	.03

* CD = complete data estimates

TABLE 1: Index Values by Parameter, *G*, and Imputation Method

		σ_0^2	σ_{01}	σ_1^2	σ_ε^2
BIAS					
	F	13.2	115.0	134.7	727.0
	p*	.015	.0001	.0001	.0000
PRECISION					
	F	11.9	30.9	255.0	95.6
	p	.021	.0002	.0000	.0003
ACCURACY					
	F	.09	10.4	38.7	8.3
	p	.92	.006	.001	.018

* The usual degrees of freedom for these tests is (2,8). Huynh & Feldt (1976) corrections were applied giving the significance probabilities shown.

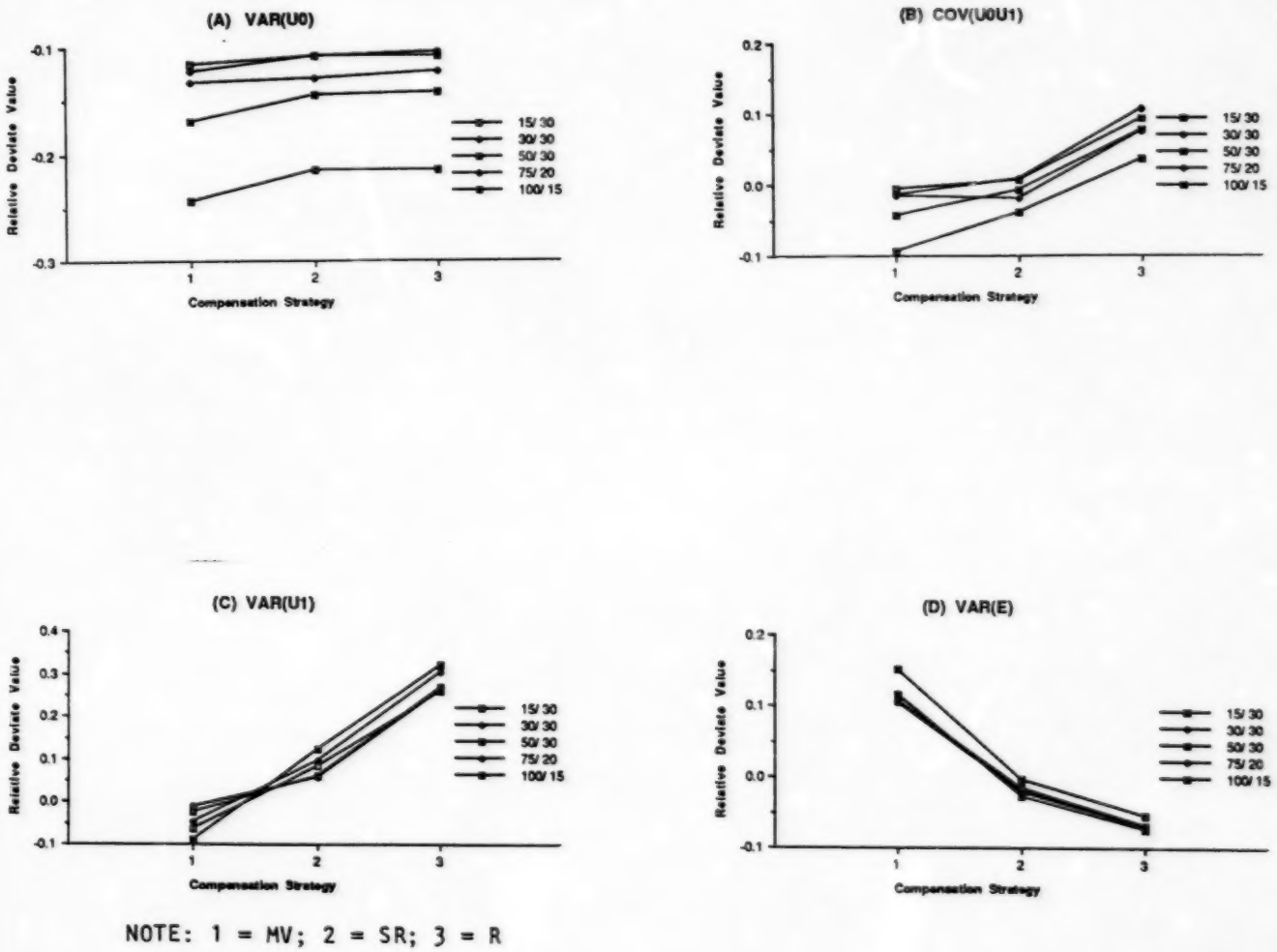
TABLE 2: F Ratios and *p* Values for ANOVA Tests of Method Effects

markedly (as expected) when the number of groups increased from 15 to 50. In the three blocks of replications in which the total number of cases was 1500, precision differences were relatively small; for each method, the precision was slightly worse when the group size was 15 than when it was 20. The ANOVA test provided some evidence of intermethod precision differences. MV imputation appeared to lead to result in a lower level of precision than did SR and R imputation.

For all imputation methods, accuracy values–displayed in Figure 3b–showed a monotonic improvement as the number of groups increased. The profiles appear to be flatter than the precision profiles, and the ANOVA test resulted in a failure to reject the null hypothesis for method effects.

4.2 Estimates of Intercept-Slope Disturbance Covariance

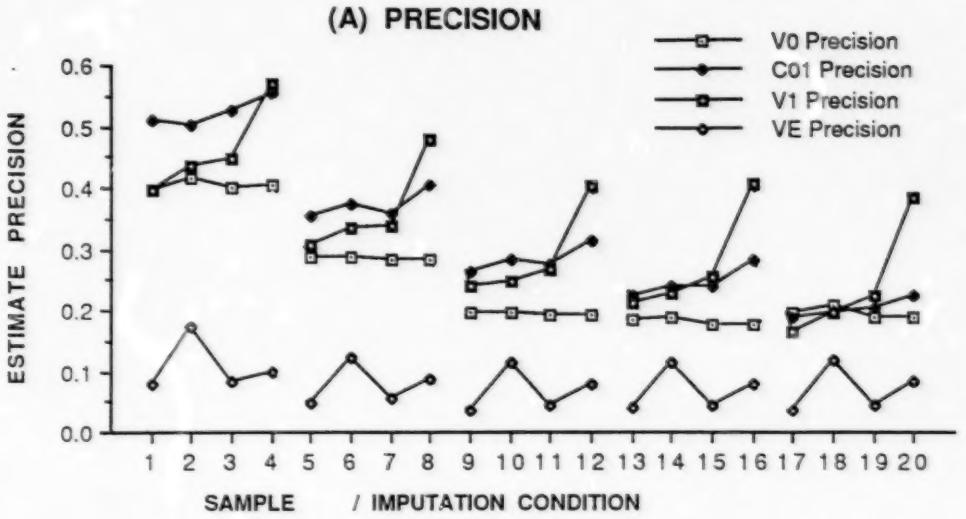
The bias in complete-data estimates was negative and smaller than 3% for all G values. These bias values improved as the number of groups increased from 15 to 75. The bias



NOTE: 1 = MV; 2 = SR; 3 = R

FIGURE 2: Bias Values for Each *G* Level and Imputation Method

under MV imputation was also always negative–the bias varied from 1% to 10%–while the bias under R imputation was always positive, ranging in magnitude from 4% to 11%. An ANOVA test of the method effect on bias was highly significant, as expected ($F_{2,8} = 115.0; p = 0.0001$). Figure 2b suggests that there may be a method by *G* interaction. This should be investigated further.



LEGEND:

1 CD G = 15/30
2 MV
3 SR
4 R
5 CD G = 30/30
6 MV
7 SR
8 R
9 CD G = 50/30
10 MV
11 SR
12 R
13 CD G = 75/20
14 MV
15 SR
16 R
17 CD G = 100/15
18 MV
19 SR
20 R

CD = complete data

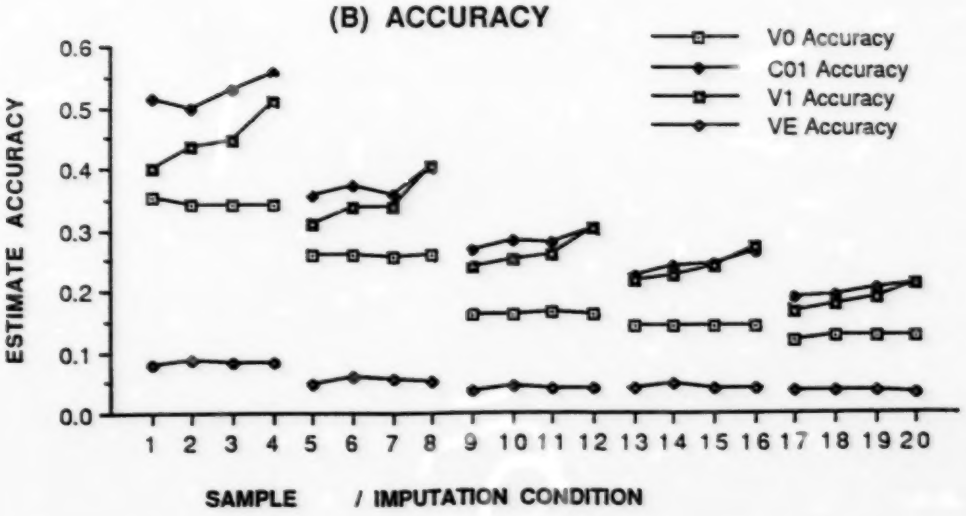


FIGURE 3: Precision and Accuracy Values by *G*/ Imputation Condition

Precision and accuracy improved monotonically, for each imputation method, as the number of groups increased. (See Figures 3a and 3b.) There appeared to be a method by G interaction with respect to both precision and accuracy. For each index, R imputation was worst, but the ranking of MV and SR depended upon the number of groups. When 50 or 75 groups were used, SR was better; otherwise, MV was superior. The ANOVA test for a method effect on precision was highly significant ($F_{2,8} = 30.9; p = 0.0002$), as was the test concerning effect on accuracy ($F_{2,8} = 10.4; p = 0.006$).

4.3 Estimates of Slope Disturbance Variance

There was no clear G-related trend for bias in estimation with complete data, but underestimation of $\sigma^2_{v_1}$ occurred for all numbers of groups. Bias ranged from 1% to 3%. Mean value imputation resulted in underestimation–see Figure 2c–but SR and R imputation lead to overestimation. The three imputation methods differed significantly in the bias they produced ($F_{2,8} = 134.7; p = 0.0001$). When absolute values of bias were examined, MV values were more satisfactory than SR values, and R values were always worst.

Figure 2c indicates the presence of G by method interaction with respect to bias. For MV, the second worst and worst levels of bias occurred when 15 and 100 groups, respectively, were used. On the other hand, bias was worst for SR and R when the number of groups was 100, and second worst with 75 groups.

With one slight anomaly, precision improved monotonically for all methods as the number of groups increased. Further, for each number of groups, estimation following MV was more precise than estimation following SR; precision associated with R was invariably worst. The method effect on precision was highly significant ($F_{2,8} = 255.0; p < 0.0001$).

The improvement of accuracy as number of groups increased was perfectly monotonic for all methods, and, with one exception, the accuracy ordering of the methods–best to worst–was MV, SR, R for all numbers of groups. The effect of method on accuracy was highly significant also ($F_{2,8} = 38.7; p = 0.001$).

4.4 Estimates of Level 1 Disturbance Variance

Estimation using complete data resulted in bias of less than 1% for all numbers of groups, and when SR imputation was used, estimates were downwardly biased by less than 3%. MV and R imputation were comaratively poor strategies, leading to overestimation with the former of 10% to 15% and underestimation by 6% to 8% with the latter. Figure 2d indicates that there was not a monotonic improvement in bias with number of groups for any method and illustrates a method by G interaction. There was a clear method effect ($F_{2,8} = 727.0; p < 0.0001$).

Precision increased monotonically with number of groups for complete-data estimation, but improvement stopped with 50 groups for each imputation method. Clear differences among the methods emerged ($F_{2,8} = 95.6; p = 0.0003$), with MV worst, and SR best for each number of groups.

Accuracy improved as the number of groups increased for complete-data estimation and for estimation following application of each imputation method. Differences among methods were comparatively small ($F_{2,8} = 8.3; p = 0.018$).

4.5 Some Inter-parameter Comparisons of Estimate Quality

Bias results in Table 1 indicate that underestimation was a problem in the determination of $\hat{\sigma}^2_{v_0}$, $\hat{\sigma}_{v_0 v_1}$, and $\hat{\sigma}^2_{v_1}$ from complete data. Estimation of the level 1 disturbance variance was unbiased when no data were missing. The problem was most severe for $\sigma^2_{v_0}$ and, overall, was about as serious for $\sigma_{v_0 v_1}$ as for $\sigma^2_{v_1}$. Mean value imputation resulted in underestimation of level 2 disturbance covariance and variances, and overestimation for level 1 disturbance variance. For disturbance variances associated with the intercept, MV imputation generally produced larger bias problems than the other methods. Estimation of $\sigma^2_{v_0}$ and $\sigma^2_{v_1}$ seemed to be most adversely affected by R imputation, and bias following SR appeared to be most severe in the determination of $\hat{\sigma}^2_{v_0}$.

Estimation of level 1 disturbance variance was always more precise and accurate than estimation of the other parameters. Figures 3a and 3b indicate that when the number of groups was 75 or fewer, the runner-up parameter with respect to estimation precision and accuracy was $\sigma^2_{v_0}$. Except when the number of groups was 15, precision following SR was about equal for $\sigma_{v_0v_1}$ and $\sigma^2_{v_1}$. The same statement could be made with respect to accuracy.

5 SUMMARY, SOME IMPLICATIONS, AND FUTURE DIRECTIONS

In this section, caveats arising from the study's limitations will be noted; some tentative data analytic advice will be offered anyway; and plans for future research will be revealed. It seems useful to begin the discussion, however, by suggesting several reasons why multilevel researchers might be concerned about mishandling missing data to put these preliminary findings into perspective,

Concerns about dealing inappropriately with missing data in multilevel analysis can be related to the substantive purposes of the research. In some studies, the magnitudes of variances among units at level 1 and at level 2 may be of substantive interest. In education, as Raudenbush and Bryk (1987, p. 241) have noted, focusing on mean differences "offers only a narrow view of the potential effects that organizational units may have on student learning... it is quite plausible that different educational processes will produce differences in the *variability* of achievement." In particular, they point out that educational programs sometimes aim explicitly at variance reduction and provide examples of this. Estimates of the relative sizes of the different levels' variances may also be important. Clearly, biased estimates of the variances can lead to incorrect conclusions, and use of compensation strategies associated with greater bias in the relevant components should be avoided.

In other studies, attention is centered on developing a model for a phenomenon, and the development process often involves testing hypotheses that differences in model coefficients across groups are real. During a preliminary stage in an example analysis from school effectiveness research discussed by Raudenbush and Bryk (1986), a test is made of the variability of the random intercept to see "whether the observed differences among schools in mean achievement level could have occurred by chance alone" (p. 7). Bias and/ or high variability in variance estimates could lead to false rejections or acceptances of null hypotheses and consequent misspecifications. Models in which coefficients should truly be regarded as random, for example, might be oversimplified if the analyst concluded on the basis of a faulty test that fixing the slopes/ intercept was acceptable.

Raudenbush and Bryk (1986) provide a third illustration of the importance of using high quality estimates of parameter variances. When level 2 explanatory variables have been entered in a model, it is sometimes of interest to assess the degree to which they account for variance in particular parameters. For this purpose, the authors suggest a coefficient of determination, R^{2*}: in a situation, for example, in which W is being used to "explain" dispersion in slope β_1:

$$R^{2*} = 1 - \frac{\mathcal{V}(\beta_1) - \mathcal{V}(\beta_1|W)}{\mathcal{V}(\beta_1)}$$

Poor estimates of the dispersions would lead to a poor sense of the usefulness of the explanatory variable.

Bearing in mind these needs for good estimates of variance components in multilevel analysis of educational and social data, one could become quite depressed after examining some of these preliminary findings–particularly those concerning large negative biases in $\hat{\sigma}^2_{v_0}$s obtained from *complete data*, and the apparent lack of improvement (sometimes) in estimation with increasing G. It may thus be comforting to think of some limitations of this small study. The unpleasant findings may then be branded as tentative, and hope for better results can be renewed.

	σ_0^2	σ_{01}	σ_1^2	σ_ϵ^2
BIAS	R / SR	SR	MV NOT R	SR NOT MV
PRECISION	R / SR	SR / MV	MV NOT R	SR NOT MV
ACCURACY	R / SR / MV	MV	MV	R / SR

FIGURE 4: Summary of Imputation Suggestions

Several shortcomings in the experiment come to mind quickly. Only one simple multilevel model was examined, and only one set of dispersion parameter true values was used in the simulations. Previous experience suggests to the researcher that if the ratio of σ_ϵ^2 to $\sigma_{v_0}^2$ is decreased, the estimation of the latter variance may be less negatively biased. A different choice for the correlation of the random slope and intercept might affect estimate quality also. Each index value in Table 1 is based on only 50 replications; rerunning the entire experiment using, say, 100 replications would provide more stable estimates of these quality indicators. Other compensation strategies–notably listwise deletion–ought to have been examined. Also, as mentioned earlier, the relationship between the effects of G and group size on estimate quality was not explored. Several "next steps" to address these and other problems will be outlined at the end of this section.

What tentative advice might be given to a researcher who is faced with 20% incomplete data (at level 1) and wants to fit a simple model like the one used here? Figure 4 outlines several suggestions concerning each random parameter, but the reader is warned that quality differences implied in the table were not tested for statistical significance (i.e., multiple comparison procedures were not used in the ANOVAs described earlier). Presence of one or more of the abbreviations MV, SR, and R in a parameter column indicate a provisional endorsement of the method with respect to the row quality. It is apparent that no one method is best or even suitable "across the board", but that stochastic regression imputation never seems very bad. The SR strategy is, unfortunately, the most complex to implement, however.

Future simulation experiments will include slope-intercept correlation, ratio of level 1 to level 2 variance, disturbance distributional form,* and percentage of missingness as independent variables. BILEVEL estimates of standard errors will be examined for bias. (To this point, a researcher's practical concerns regarding estimated SEs for comparing models have not been addressed.) Further down the road, nonrandom missingness will be examined in the context of fitting the present model, and level 2 missingness (and compensation strategies) will be examined with models that have two group level explanatory variables.

---

* A multivariate skewed distribution will be used.

References

Aitkin, M., & Longford, N. (1986). Statistical modelling issues in school effectiveness studies. *Journal of the Royal Statistical Society A, 149*, 1-43.

Basilevsky, A., Hum, D., Sabourin, D., & Anderson, A. (1985). Missing data estimators in the general linear model: An evaluation of simulated data as an experimental design. *Communications in Statistics--Simulation and Computation, 14*, 371-394.

Beale, E., & Little, R. (1975). Missing values in multivariate analysis. *Journal of the Royal Statistical Society B, 37*, 377-395.

Bryk, A., & Raudenbush, S. (1987). Application of hierarchical linear models to assessing change. *Psychological Bulletin, 101*, 147-158.

Bryk, A., Raudenbush, S., Seltzer, M., & Congdon, R. (1986). An introduction to HLM: Computer program and users' guide. Unpublished manuscript.

De Leeuw, J., & Kreft, I. (1986). Random coefficient models for multilevel analysis. *Journal of Educaticnal Statistics, 11*, 57-85.

Dempster, A., Laird, N., & Rubin, D. (1977). Maximum likelihood for incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society B, 39*, 1-38.

Goldstein, H. (1986). Multilevel mixed linear model analysis using iterative generalized least squares. *Biometrika, 73*, 43-56.

Goldstein, H. (1987). *Multilevel models in educational and social research*. London: Griffin.

Goldstein, H., & McDonald, R. (in press). A general model for the analysis of multilevel data. *Psychometrika*.

Goldstein, H., & Silver, R. (in press). Multilevel and multivariate models in survey analysis. In D. Holt, T. Smith, & C. Skinner (Eds.). *Analysis of data from complex surveys*.

Hanushek, E. (1974). Efficient estimators for regressing regression coefficients. *American Statistician, 28*, 66-67.

Huynh, H., & Feldt, L. (1976). Estimation of the Box correction for degrees of freedom from sample data in randomized block and split-plot designs. *Journal of Educational Statistics, 1*, 69-82.

IMSL. (1980). *User's manual: IMSL library*. Houston, TX: Author.

Judge, G., Griffiths, W., Hill, R., Lutkepohl, H., & Lee, T. (1985). *The theory and practice of econometrics* (2nd ed.). New York: Wiley.

Kalton, G. (1983). *Compensating for missing survey data*. Ann Arbor: University of Michigan, Institute for Social Research.

Little, R. (1979). Maximum likelihood inference for multiple regression with missing values: a simulation study. *Journal of the Royal Statistical Society B, 41*, 76-87.

Little, R., & Rubin, D. (1987). *Statistical analysis with missing data*. New York: Wiley.

Longford, N. (1987). A fast scoring algorithm for maximum likelihood estimation in unbalanced mixed models with nested random effects. *Biometrika, 74,* 817-827.

Marini, M., Olsen, A., & Rubin, D. (1980). Maximum-likelihood estimation in panel studies with missing data. In K. Schuessler (Ed.). *Sociological Methodology 1980* (pp. 314-357). San Francisco: Jossey-Bass.

Raudenbush, S., & Bryk, A. (1986). A hierarchical model for studying school effects. *Sociology of Education, 59,* 1-17.

Raudenbush, S., & Bryk, A. (1987). Examining correlates of diversity. *Journal of Educational Statistics, 12,* 241-269.

Rubin, D. (1987). *Multiple imputation for nonresponse in surveys.* New York: Wiley.

Santos, R. (1981). *Effects of imputation on complex statistics.* Ann Arbor: University of Michigan, Survey Research Center.

Searle, S. (1987). *Linear models for unbalanced data.* New York: Wiley.

Tate, R., & Wongbundhit, Y. (1983). Random versus nonrandom coefficient models in multilevel analysis. *Journal of Educational Statistics, 8,* 103-120.

van den Eeden, P., & Huttner, H. (Eds.). (1982). Trend report: Multi-level research. *Current Sociology, 30.*